J. Pestic. Sci., **31**(3), 296–299 (2006)

Commentary

(Special Topic)

Overview of KEGG applications to omics-related research

Kiyoko F. AOKI-KINOSHITA*

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 600–0011, Japan

(Received March 9, 2006)

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a bioinformatics resource for analyzing cells and organisms from not only the genomic perspective but also a high-level perspective, integrating together genomic, chemical and network information.[1)] Accessible from http://www.genome.jp/, it basically consists of four databases: PATHWAY, GENES, LIGAND and BRITE. The KEGG PATHWAY database provides pathway diagrams, represented as networks of interactions that occur in the cell. These can be viewed according to organism or as generic "reference" maps. KEGG GENES is the collection of genes that are found in the complete genomes that are registered in KEGG. It serves as the repository of genes linked from the pathway diagrams. KEGG LIGAND is a database of compounds, glycans, reactions and enzymes.[2)] Finally, KEGG BRITE contains the KEGG Orthology, or KO, which is a manually curated identification system of gene orthologs. It also contains classifications of chemical compounds and enzymatic reactions. KO has become an indispensable tool for the functional annotation of new genomes, and it plays a key part in the KAAS (KEGG Automatic Annotation Server) tool.

Overview of KEGG

1. *KEGG PATHWAY*

The KEGG PATHWAY Database is a hierarchically organized collection of diagrams, listed in Table 1. Each diagram is a reference pathway that corresponds to a known network of functional significance. Each pathway map can be viewed according to organism, where species-specific maps are reference maps colored with the genes in the given organism.

An example of a pathway is given in Fig. 3 of the Review by Kadowaki *et al.* in this issue: flavonoid biosynthesis in *Arabidopsis thaliana*. In general, rectangles on a pathway map represent gene products, usually proteins. Small circles represent other types of molecules, such as chemical compounds. Large ovals containing pathway titles are linked to other pathway maps, and clusters of rectangles represent complexes. In the figure, note the shaded boxes, which indicate those genes in the pathway that can be found in the selected *Arabidopsis thaliana* species.

2. *KEGG GENES*

The KEGG GENES database originated from the gene data in GenBank, and it has since been augmented in KEGG with annotation done independently based on the literature. KEGG GENES can be considered a collection of "genome databases" where one genome is a database containing entries of genes. Note that the KEGG nomenclature specifies a three-letter code for each genome, such as *hsa* for *Homo sapiens*. The list of all organisms in KEGG and their three-letter codes can be seen at http://www.genome.jp/kegg/catalog/org_list.html.

Each gene entry stores all information related to the gene including amino acid and DNA sequence information, pathways in which it can be found, motifs found in its sequence, and links to other databases such as UniProt and NCBI. The KO ortholog group (see KO section) is also specified in this entry such that other orthologons genes in other organisms in the same pathways can be directly accessed.

3. *KEGG LIGAND*

KEGG LIGAND originally consisted of four main sub-databases: COMPOUND, GLYCAN,[2)] REACTION and ENZYME.[3)] Recently, two new sub-databases DRUG and RPAIR have also been added. COMPOUND is a collection of chemical structures, mostly known metabolic compounds and some pharmaceutical and environmental compounds. GLYCAN consists of carbohydrate structures, REACTION is a database of reaction formulas for enzymatic reactions, and ENZYME is a database of enzyme nomenclatures. DRUG is complementary to the COMPOUND database, containing the

* To whom correspondence should be addressed.
E-mail: kiyoko@kuicr.kyoto-u.ac.jp

Table 1. The KEGG PATHWAY Database is organized hierarchically

Level 1	Level 2
Metabolism	Carbohydrate metabolism
	Energy metabolism
	Lipid metabolism
	Nucleotide metabolism
	Amino acid metabolism
	Metabolism of other amino acids
	Glycan biosynthesis and metabolism
	Biosynthesis of polyketides and nonribosomal peptides
	Metabolism of cofactors and vitamins
	Biosynthesis of secondary metabolites
	Xenobiotics biodegradation and metabolism
	Enzyme families (Cytochrome P450)
Genetic Information Processing	Transcription
	Translation
	Sorting and degradation
	Replication and repair
Environmental Information Processing	Membrane transport
	Signal transduction
	Signaling molecules and interaction
Cellular Processes	Cell motility
	Cell growth and death
	Cell communication
	Immune system
	Nervous system
	Development
	Behavior
Human Diseases	Neurodegenerative disorders
	Infectious diseases
	Metabolic disorders
Drug Development	Chronology of drug development
	Target based structure classification
	Skeleton based structure classification

structures of compounds used as drugs, and the RPAIR database contains all patterns of reactant pair transformations in the REACTION database.

All chemical structures in the KEGG COMPOUND database are manually entered, computationally verified and continuously updated. It contains over 11,000 entries, with each entry ID beginning with the letter C. The KEGG GLYCAN database contains carbohydrate sugar chains, or glycans, whose entry IDs are prefixed by the letter G. Most of the over 10,000 entries in GLYCAN originated from the CarbBank database, which has been cleaned up by combining duplicates and making the content consistent. Pathway diagrams on the metabolism of complex carbohydrates and complex lipids are linked from these glycan entries.

The KEGG REACTION database contains over 6000 reaction formulas for enzymatic reactions. Each entry ID is prefixed by the letter R, representing a unique reaction of sets of reactants and products. This contrasts with EC numbers, where each EC number may correspond to multiple reaction formulas. Each entry is linked with the entries in the COM-

POUND or GLYCAN databases that correspond to the molecules that participate in the reaction.

The DRUG database currently contains almost 3000 compounds, with entry IDs beginning with the letter D. Many are classified according to therapeutic applications, and of course any available links such as to PubMed and CAS IDs are also linked. These structures are also linked with the Drug Structure Maps under the Drug Development category of KEGG PATHWAY.

The RPAIR database stores patterns of transformations that may occur between two reactants in a single reaction. This database is still under development, but reactions are linked to its related entries.

4. BRITE database

The BRITE database contains hierarchically organized categories of data, including the KO (KEGG Orthology) hierarchy and drug classifications. KO is a manually curated pathway-based classification of ortholog groups. Each KO ID begins with the letter K. Using the highly-confident Smith-Waterman algorithm to align all pairs of sequences in KEGG, similarity scores are calculated for all pairs of genes. From these scores, ortholog groups are classified according to pathway and provided as a hierarchical classification which allows highly-similar genes across organisms to be compared on the pathways. The entire KO hierarchy can be viewed using the KO link on the main BRITE page at http://www.genome.jp/kegg/brite.html. BRITE also contains classifications of drug compounds based on therapeutic categories as well as classifications of compounds and protein families, to name a few. It is also the basis on which the KAAS tool for automatic KO assignments function, described later.

Using KEGG for -Omics Research

KEGG has a number of useful applications for research. Pathways may be colored and analyzed based on a specified organism or a user-defined list of genes. They are also provided in XML format such that they can be downloaded and analyzed locally. BLAST and FASTA searches can be run on the genes and genome databases of KEGG, and microarray expression analysis tools are also available. Chemical compounds can be analyzed in a variety of ways with the compound comparison tools available. Finally, the KAAS (KEGG Automatic Annotation Server) tool utilizes the KO classifications to predict functional annotations of gene sequences.

1. Pathway coloring

The pathways can be colored according to a specific organism such that those genes that are found in the given organism are highlighted. Furthermore, given a list of gene names, pathways can be customized to display specific genes with specific colors. This tool is available under the "Color objects in KEGG pathways" option under the KEGG Table of Contents.

2. Chemical compound search of the LIGAND database

The LIGAND database has several tools for analyzing chemical compounds in terms of similarity comparison and reaction prediction. Similarity comparison involves the specification of a simple compound or glycan structure and the retrieval of all similar structures from the COMPOUND and GLYCAN databases. The easiest method is to download the KegDraw application from http://www.genome.jp/download/ (which also contains a link to download the microarray data analysis software KegArray), and using the "Search Similar Structures" option under the "Tools" menu. Alternatively, a web-based interface for querying the COMPOUND and GLYCAN databases is also available. All queries, whether from KegDraw or over the web, use the various compound comparison tools provided in KEGG: SIMCOMP,[4] SUBCOMP and KCaM.[5]

3. Reaction prediction tools

Another tool, which takes two chemical compounds as input, predicts the possible series of reactions that can take place in order to begin with one compound and end with the other. This is the PathComp tool. Specific pathways can be specified to search for the reaction paths, so that the organisms containing specific compounds can be retrieved, for example. Another tool is the reaction prediction tool called e-zyme,[6] which takes at least one pair of compounds and returns all possible reactions involving them. These results are all linked with PATHWAY such that a listing of all pathways involving given compound structures can be retrieved.

4. KAAS: KEGG Automatic Annotation Server

The KAAS tool at http://www.genome.jp/kegg/kaas/ provides a web-based service for annotating a set of (unknown) genes. By inputting a set of sequences, they are compared with the KO-annotated genes of KEGG, and the annotation for each input gene is returned. Thus new completely sequenced genomes can be quickly annotated with functional information using KAAS. Pathways involving the given genes would allow users to automatically generate pathways, consequently making it possible to see which related organisms contain the given genes (using KO). This tool is becoming increasingly useful for biologists as more new genomes become completely sequenced. Note that gene sets from entire genomes is not required to use KAAS; sets of ESTs can also be annotated automatically as well.

Conclusion

In summary, KEGG offers researchers many tools for -omics research from pathway prediction to genome comparison and chemical compound comparison and analysis. Further details regarding the usage of these tools and more have been published.[7] Feedback is also always welcome as KEGG is continually being improved based on user feedback. The goal of KEGG is to provide a comprehensive resource for analyzing the complex biological world not only within organisms, but

also interacting with the environment, in the hopes of advancing -omics research.

References

1) M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa: *Nucleic Acids Res.* **34**, D354–D357 (2006).
2) K. Hashimoto, S. Goto, S. Kawano, K. F. Aoki-Kinoshita, N. Ueda, M. Hamajima, T. Kawasaki and M. Kanehisa: *Glycobiol.* **16**, 63R–70R (2006).
3) S. Goto, Y. Okuno, M. Hattori, T. Nishioka and M. Kanehisa: *Nucleic Acids Res.* **30**, 402–404 (2002).
4) M. Hattori, Y. Okuno, S. Goto and M. Kanehisa: *J. Am. Chem. Soc.* **125**, 11853–11865 (2003).
5) K. F. Aoki, A. Yamaguchi, N. Ueda, T. Akutsu, H. Mamitsuka, S. Goto and M. Kanehisa: *Nucleic Acids Res.* **32**, W267–W272 (2004).
6) M. Kotera, Y. Okuno, M. Hattori, S. Goto and M. Kanehisa: *J. Am. Chem. Soc.* **126**, 16487–16498 (2004).
7) K. F. Aoki and M. Kanehisa: "Current Protocols in Bioinformatics," Vol. 1, ed. by L. Miranker, John Wiley & Sons, Ltd., Chap. 12, 2005.